

# Teat Pose Estimation via RGBD Segmentation for Automated Milking

Nicolas Borla<sup>1</sup>, Fabian Kuster<sup>1</sup>, Jonas Langenegger<sup>1</sup>, Juan Ribera<sup>2</sup>, Marcel Honegger<sup>1</sup>, Giovanni Toffetti<sup>2</sup>  
Zurich University of Applied Sciences (ZHAW)

**Abstract**—We present initial results in the development of a novel robot using RGBD cameras, image segmentation, and a simple teat pose estimation algorithm for automated milking. We relate on the analysis of the accuracy of different commercial RGBD cameras in realistic conditions. Although preliminary, our initial implementation shows that 2D image segmentation combined with point cloud processing can achieve repeatable millimeter-scale precision in estimating (synthetic) teat tip positions and cup attachment approach. The solution is also applicable in a cloud robotics setup, with GPU-based segmentation executed on an edge device or cloud.

## I. INTRODUCTION

Milking robots have been in use for almost 30 years, after the first systems were installed in the Netherlands in 1992 [1]. Today, the main suppliers of milking robots are Lely, DeLaval and GEA Farm Technologies. Their systems are typically used on large dairy farms and are optimized for milking cows up to 3 times per day, and around the clock. However, in many European countries such as Switzerland, dairy farms typically have less than 60 cows and they are usually milked only twice, once in the morning and once in the evening, while they are left to graze during the day. This use case requires milking robots that occupy less space, so that it is easier to install several systems in parallel in existing barns, and it requires the robots to milk cows as efficiently as possible, because the time slots for milking are far shorter. For this reason, we started a research project to design a new generation of milking robot, using the latest technologies to reduce the space required for the milking robot manipulator, and to reduce the time required to milk cows.

While most existing milking robots only offer 3 degrees of freedom (DoF) [2], this new manipulator offers 5 DoF, so that the milking cups can be positioned in all Cartesian coordinates, and their orientation aligned to the orientation of the teats. This allows to more reliably place and attach the milking cups to the teats.

Attaching milking cups to teats often takes more than a minute with existing milking robots [3]. Before attaching cups, these robots often scan the teats with laser scanners to detect the teats positions, then the manipulator moves the cups to the teats for attachment. Attachments fail in 5 to 10% of all cases [3], which requires another scan of the

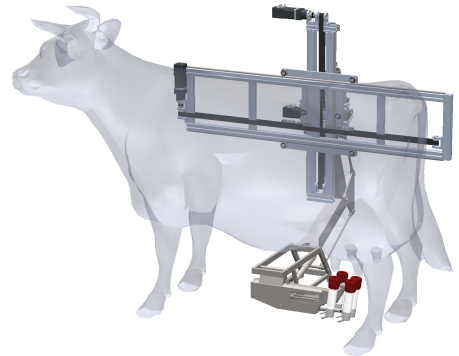


Fig. 1. Kinematic model of the milking robot

teats, for another trial to attach the cups.

With a new approach to detect the teat poses with RGBD cameras in real-time, attaching the cups to teats does not require a preceding, time-consuming scanning procedure. Instead, the teat positions and orientations are detected while the robot manipulator is moving the cups to the estimated position of the teats, and is adjusting the motion to the detected positions with each measurement. This will reduce the time needed to attach the cups to the teats, and increase the reliability of the attachment process significantly.

## II. RELATED WORK

In the following subsections we relate on the state of the art of commercial automated milking robots, published research in the field, and more general algorithmic approaches that can be applied to the problem at hand.

### A. Commercial Milking Robots

Five major international commercial milking robot suppliers provide their services worldwide: Lely, DeLaval, GEA, SAC, and Lemmer-Fullwood. Apart from them, there are several smaller (typically national) suppliers. A complete discussion of the pros and cons of each solution is beyond the scope of this paper, moreover there are several specialized publications offering such comparisons online<sup>1</sup>.

Overall, the major drawbacks common to all current commercial solutions are:

- High cost per milking robot unit;
- Intense robot usage to amortize investment cost;
- High cost of replacement parts and materials;

This research project is supported by Innosuisse, the Swiss Innovation Agency, with project number 40368.1 IP-ENG.

<sup>1</sup> The authors are with the Institute of Mechatronic Systems (IMS), Technikumstrasse 5, 8401 Winterthur, Switzerland

<sup>2</sup> The authors are with the Institute of Applied Information Technology (InIT), Obere Kirchgasse 2, 8400 Winterthur, Switzerland

<sup>1</sup>E.g., <https://www.melkroboter.net> (in German)

- Limited or no support for continuous learning / adaptation to cow udder morphology (i.e., personalization)

### B. Automated Milking Literature

Many of the current state of the art solutions rely on laser scanner technology to detect teats and estimate their pose. A 2D laser scanner implies a scanning procedure, moving the sensor to different heights to achieve a 3D measurement. A significant drawback of this design is that the measurement cannot be performed in real-time. If the cow moves, a new measurement procedure is needed before the cups can be attached. Relying purely on depth information, laser technology may fail in correct teat identification, therefore manipulating the suction cups in the wrong direction. For this reason, [4] proposes a fast and reliable solution to the problem using Time of Flight (ToF), RGBD and Thermal Imaging. The study from [5] for vision systems for livestock reports that RGB-D technologies are preferable to ToF cameras.

Similarly, [6] takes a stand against the limitations of laser assisted edge detection technologies, which cannot differentiate between a healthy and a diseased teat. They propose two alternatives to the task: a Haar-cascade classifier and a YOLO classifier for cow teats. Both approaches work on real time but lack reliable accuracy.

In [7] several references are given to teat pose estimation algorithms applied by commercial milking robots. However, the authors state that their method to identify teat tip positions from low resolution 3D-ToF camera videos is superior to all previously reported ones. The method is based on edge detection on the depth image combined with matching U-shaped templates. To account for teat size and distance from camera, resized U-shapes are applied for correlation. In order to account for non-vertical teats, PCA (principal component analysis) is used to obtain rotational invariant teats. The proposed solution requires limited computation and achieves teat pose estimation at 4 to 8 FPS. Validation results show “90% of the frames being successfully tracked” on 15 videos.

The work in [8], discusses the application of 3D vision technologies to precision livestock farming, including in automated milking, and concludes that at time of publication (2019) 3D deep learning solutions were not yet applicable due to a lack of sufficient training data, a problem common to all 3D deep learning computer vision applications.

Finally, in [9], the authors use a 3D-ToF camera to collect both RGB images and a point cloud. They process the point cloud applying the k-nearest algorithm for segmentation, but such method cannot distinguish the udder from the teats, resulting in imprecise segmentation. To counter this problem, their method relies on assumptions on the camera position w.r.t. the teat for teat detection. Still, this prevents them from correctly identifying teats on real cows where udder morphology is highly variant.

### C. Object Detection, Pose Estimation, Grasping

As reported in [10], “the ability for robots and computers to see and understand the environment is becoming a

burgeoning field, needed in autonomous vehicles, augmented reality, drones, facial recognition, and robotic helpers”. Since the rise of the CNN [11] deep learning based methods for image classification have reached state of the art performance for 2D detection. Nevertheless, 3D scene interpretation methods continue to struggle because of 1) the lack of publicly available RGB-D data sets [12] and 2) the not yet widespread adoption of depth cameras compared to 2D ones [10].

Several algorithms have been developed for automated pose estimation and / or grasp generation of objects in literature, for instance [13], [14], and [15]. The approaches above address the general problem of grasping and manipulating unknown objects, however they cannot directly be applied to our specific manipulation task (i.e., teat attachment and successful pumping). One possible approach that goes in the direction of generalizing object classes and their manipulation is [16] which uses semantic 3D keypoints for object representation and enables the specification for robot action planning and grasping with centimeter level precision. Our initial work done in this paper is a needed step to try to apply that kind of approach to automated milking.

## III. REQUIREMENTS

The initial requirements from the project specification for the teat pose estimation algorithm prototype were:

- Pose estimation must be performed continuously during the movement of the robotic arm;
- Maximum estimation time of all teat poses of 10 seconds (this includes any arm movement required to reduce uncertainty);
- Recognition of the correct pose (within an error of 0.5 centimeters) in at least 90% of teats.

Apart from these formal requirements coming from the project contract, further requirements for the solution stemmed from the fact that the robot shape and kinematics had to be designed, hence further requirements for the algorithm are:

- No assumptions should be made about camera(s) positions and orientations w.r.t. the udder;
- Occlusions have to be expected and taken into account;
- No assumptions should be made about teat number<sup>2</sup>, positions, and orientations (cow’s udder morphology);

No initial requirements were given with respect to the architecture and cost of the compute unit (or GPU) to be used for the implementation.

## IV. SOLUTION ARCHITECTURE

Given the requirements from the previous section, and our analysis of the state of the art from Section II, we oriented ourselves in choosing a solution that would allow us to minimize assumptions (e.g., on poses / frames) and at the same time account for natural variation (i.e., changing udder morphology, teat colors, light conditions). This lead us to restrict the space of solutions towards a combination

<sup>2</sup>Not all cows have exactly 4 teats [9]

of Neural Networks (NNs) based on Deep Learning (DL) with pose estimation from point clouds (PCLs).

In this respect, our review of the literature of DL solutions applied to 3D convinced us that, at that specific moment, we could not leverage any existing 3D DL technology for the project, be it for reasons of prediction performance (both in terms of rate and accuracy) or training cost (including training data set labelling).

Upon this reasoning, we decided to build on existing mature DL technologies to identify teats from 2D color images. The rationale here is that 2D DL allows us to minimize false positives in recognition while accounting for natural variation of morphology, colors, and light conditions. Here, discounting the different NN models, the main decision to make was whether to use multi-object identification (i.e., bounding boxes around teats) or multi-object segmentation (i.e., a pixel mask for all pixels belonging to each teat). We opted for the second alternative which, albeit slower (e.g., with Mask-RCNN), allowed us to have a more precise mask closely matching the shape of the teats as seen in 2D.

To bridge the gap from 2D to 3D, we borrowed the idea from [17] to project the mask stemming from the 2D teat identification step into the point cloud with a frustum to “carve out” teats in three dimensions. Then, for each 3D teat candidate, a combination of clustering, PCA, and surface normals algorithms can be used to estimate the orientation of a teat, identify the tip, and compute the its 6 DOF pose in 3D space.

The resulting high level functioning of our solution is depicted in Figure 2.

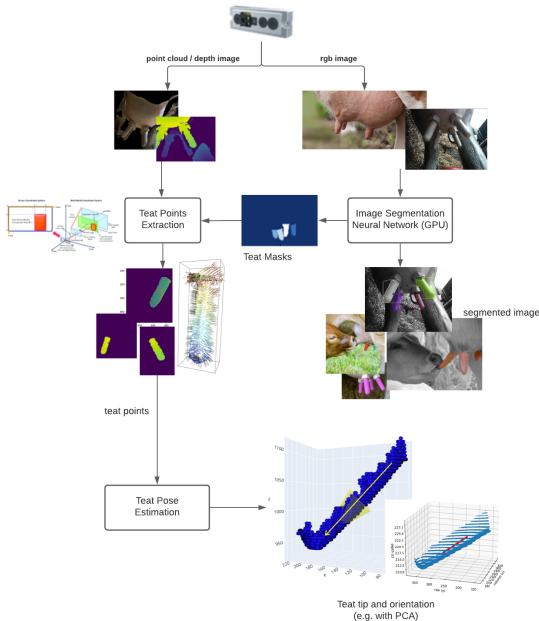


Fig. 2. Main functional blocks of our teat pose estimation solution

## V. IMPLEMENTATION

The implementation of the solution was distributed across the labs participating in the project based on expertise.

ICCLab (InIT) focused on the teat pose estimation algorithm, while the IMS took responsibility of all the robotic aspects, from the evaluation of different cameras for the task at hand, to their calibration and correction of errors, to the design of the final robot and the programming of the arm control logic.

Given the distributed nature of the project and the possibility to apply cloud robotics solutions to the final product, we decided to implement the software stack for the project as a distributed system from the get go. In particular, we used containerization and a multi-master ROS design to isolate the different versions of operating systems, ROS, and libraries that were needed for the different components of the project.

The overall component architecture of the final implementation is depicted in Figure 3.

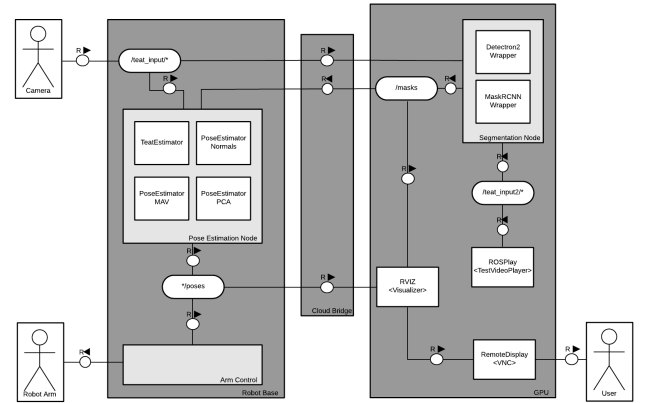


Fig. 3. FMC architectural diagram of our teat pose estimation solution

We opted for using three different ROS nodes each implementing a subset of the functionalities:

- the *Pose Estimation node* is the interface to the sensor data and the robot. It receives the (time synchronized) messages from the camera sensors (i.e., RGB image and point cloud), forwards the RGB image to the NN, awaits for the teat masks to be detected, and uses the masks to publish the detected teat poses;
- the *Segmentation node* hosts the NN that performs the multi-object image segmentation and publishes the detected masks;
- the *Arm Control node* receives multiple messages about estimated teat poses and performs arm movement planning if a configurable number of teat pose estimates is consistent over time.

In the following subsections we relate on the implementation of each of the nodes.

### A. Pose Estimation Node (*find.teat.poses*)

This is the node that connects the robot to the segmentation neural network node. A synchronized message filter receives both the point cloud and the `rgbImage` published at the same time instant. The node saves the point cloud in memory and forwards the `rgbImage` to be processed by the NN.

After segmentation, the NN publishes the segmented image for visualization and the "masks" resulting from segmentation. We use a project specific ROS message format to reduce the amount of data that is passed back from the NN to represent the masks.

Upon reception of the masks the node `find_teat_poses` uses the 2D mask contour to extract the corresponding 3D points from the point cloud and to estimate the position and orientation of all visible teats. Teat points are extracted from the point cloud by first applying a voxelization step, then projecting the rays matching each teat masks contour in 3D space and removing anything outside of the generated frustum.

Finally a set of pose estimation algorithms are applied to estimate teat tip positions and the required orientation of the teat cup for a successful attachment. All teats poses are published and visualized with a marker as in Figure 4.

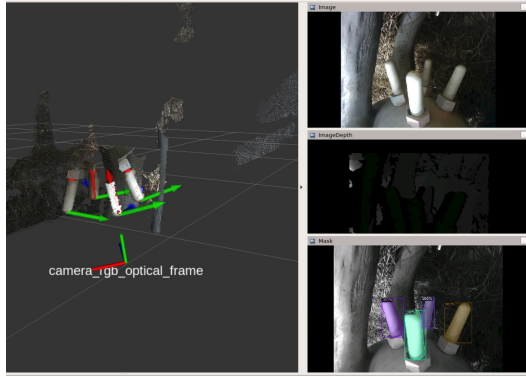


Fig. 4. Visualization of teat pose markers and teat segmentation

**Teat Tip Pose Estimation Algorithms:** With the 3D points for each teat as input, different algorithms can be applied to understand the location of the teat tip and the required orientation of a teat cup to perform an attachment. We relied on two simple implementations from geometric principles: principal component analysis (PCA) and using surface normals.

With the PCA algorithm, the underlying assumption is that cow teats have a generic cylindrical shape and are longer than wider, hence running PCA on the 3D points of a teat would yield the teat "cylinder axis" as the main principal component. Given the fact that the teat axis can be represented by two vectors with opposite orientation, we use the camera position to identify the vector direction for cup attachment (i.e., upwards rather than downwards) hence the teat tip.

The surface normals algorithm is based on a different idea to identify a teat axis. That is, if a cow teat is approximately cylindrical, the vectors that are orthogonal to its surface (i.e., "surface normals") will also be normal to the teat axis. Hence, the teat axis direction can be estimated by finding the vector that minimizes the sum of the dot products with the vector itself for each surface normal. As in the case of the PCA algorithm, a further step to correctly identify the vector orientation for teat attachment has to be performed.

## B. Segmentation Node (Neural Network)

In the course of the project we experimented with different publicly available neural network implementations to perform either object detection or multi-object segmentation.

In the end we built our prototype based at on two implementations of the MaskRCNN paper: Matterport's MaskRCNN<sup>3</sup> and Facebook Research's Detectron2 implementation of MaskRCNN<sup>4</sup>.

Both implementations are highly configurable and allowed us to use 640x480 pixel images as input wrapping the invocation of their inference functionality in a simple ROS topic subscriber callback handler.

Benchmarks<sup>5</sup> of both implementations show clear differences in the Average Precision (AP) between them showing Detectron2 having better accuracy. Moreover, benchmarks show the implementations from matterport have a 4x slower throughput (imgs/sec) compared to Detectron2<sup>6</sup>. These drawbacks and the generally better performance of the Detectron2 implementation led to it being the favorite for the segmentation task.

## VI. RESULTS

In this section we relate on the methodology and results obtained in evaluating different commercial cameras and implementing a first prototype of a complete teat pose estimation and attachment solution.

### A. 3D sensor

An essential requirement for the 3D sensor mentioned in the previous sections is that pose estimation must be performed on the fly during arm (and cow) movement. Therefore, no sensor which needs a scanning procedure is suitable. We selected five 3D cameras among the newest models available on the market and carried out an accurate evaluation of their performance to choose the most suitable for our task. These cameras are produced by different manufacturers and use various measurement technologies:

- Orbbec Astra Embedded S
- Orbbec Astra Stereo S U3
- PMD/Infineon CamBoard Pico Flexx
- Intel RealSense SR305
- Intel RealSense D435

To estimate distances, the two cameras of Orbbec and the Intel D435 use active IR stereo vision, the Pico Flexx uses an IR Time-of-Flight sensor, and the Intel SR305 uses structured light. All cameras except the Pico Flexx integrates an RGB camera, meaning that they provide data in the form of 3D point clouds and 2D colour images.

To evaluate the five cameras, we used a test setup already available at IMS for the test of general purpose 3D sensors [18]. This setup consists of a steel plate with plastic stops

<sup>3</sup>[https://github.com/matterport/Mask\\_RCNN](https://github.com/matterport/Mask_RCNN)

<sup>4</sup><https://github.com/facebookresearch/detectron2>

<sup>5</sup><https://github.com/facebookresearch/maskrcnn-benchmark/issues/449>

<sup>6</sup><https://detectron2.readthedocs.io/en/latest/notes/benchmarks.html>



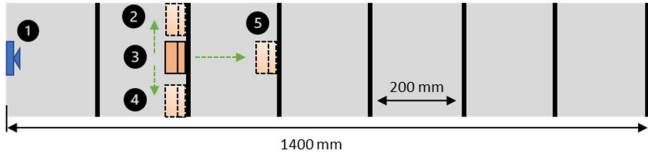


Fig. 5. Test setup used to evaluate the performance of different cameras: (1) 3D camera, (2) test object placed in the left track, (3) test object placed in the central track, (4) test object placed in the right track, (5) the distance is measured placing the test object further away from the camera (every 200mm)

every 200mm up to a distance of 1.4m with three tracks: left, centre and right. The test object is a 3D printed L-form with a surface of 100x150mm (the surface is orthogonal to the camera). The test object is placed every 200mm, and the distance is measured by averaging the points measured on the surface. Figure 5 illustrates the concept of the test setup.

The absolute accuracy of the distance measurement for all cameras is evaluated under different conditions using the test setup. The different situation evaluated are:

- Different light condition (direct light from headlamp, room light or night)
- Different colour of the test object (White, Black or Pink like the teats of the cow)
- Lateral shift (change track)
- Influence of a 5mm glass panel in front of the camera (to simulate the sealing needed to work outdoor)
- Variance (repeatability of the measurement)
- Influence of the camera resolution (for the cameras which offer a configurable resolution)

The results of the test are shown in figure 6. In this chart, only the maximal error for distances up to 1m is considered for all different conditions and cameras. The reasons are that the robot's working range under the cow is limited to 1m in the mechanical requirements and that it makes it easier to show the accuracy of the cameras and the influence of the test conditions.

As a general behaviour, all cameras show an error in absolute accuracy that increases approximately quadratically with distance. The maximum error at 1m distance, shown in figure 6, also reflects how fast the error grows for each camera. The variance is not illustrated because under steady conditions it is less than 0.2mm for all cameras. It is not the repeated measurement that introduces a relevant error in the measurement but rather a change in the measurement conditions.

It can be noted that the PicoFlexx, the Intel SR305 and the Orbbec stereo have similar performances and overall are better than the other two cameras. PicoFlexx uses Time of Flight technology and seems to be more sensitive to different working conditions. Moreover, this camera has no RGB sensor included. The Intel SR305 is much larger than The Orbbec Stereo (at least twice the volume) and less accurate. Therefore, we chose the Orbbec stereo for our implementation. This camera performs overall better than all other sensors tested, has an RGB camera already incorporated, and according to the manufacturer is specifically

developed to work in a multi-camera setup. A setup with multiple cameras could be interesting for the milking robot, and therefore, the same tests were repeated using two Orbbec stereo pointing at the same object. The results showed no interference between the cameras and the same result as the setup with only one camera.

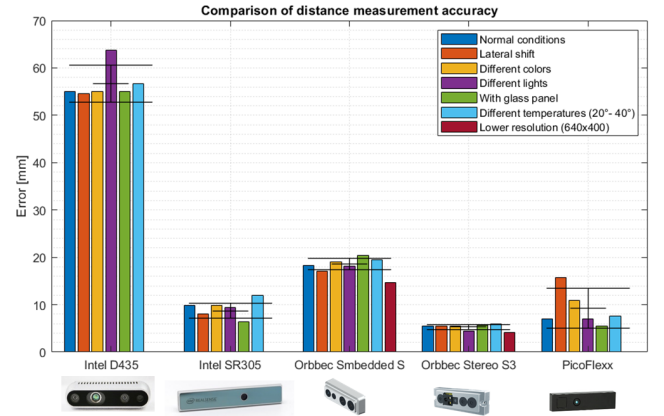


Fig. 6. Maximal distance measurement error for objects up to 1m under diverse conditions

### B. Pose Estimation Accuracy

In the current first phase of the project, experiments were conducted with a dummy cow under laboratory conditions. This framework implies an indoor environment, varying (low) light conditions, and no varying udder geometry. We used the UR10e robot as a manipulator with the Orbbec Astra Stereo S U3 camera and a single teat cup mounted on the robot flange. An Ubuntu computer was selected as a local controller with the ROS framework for software development. The 2D camera image is sent over wifi to a separate virtual machine in our local cloud computing cluster equipped with a Nvidia Tesla T4 graphic card for teat detection. Upon detection, teat masks are used to predict teat poses on the robot. To validate the four calculated teat positions, we moved the robot with the attached teat cup to each teat of the udder one after the other.

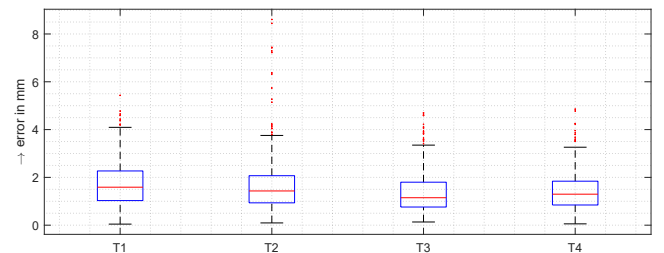


Fig. 7. Teat pose error deviation for each teat (789 measurement)

The first experimental question we ask concerns the accuracy in estimation of the pose of the teat tip and its repeatability. To evaluate this, we kept the cow model in the same position and we precisely measured teat tip poses

for each teat (T1-T4) to provide a ground truth. Then we ran 789 teat position estimation cycles under varying light conditions and initial arm positions, measured the error of the pose estimation w.r.t. the ground truth obtaining the results in Figure 7 and Table I.

in mm	mean	std
T1	1.7	0.9
T2	1.6	1.0
T3	1.4	0.8
T4	1.4	0.7

TABLE I

MEAN AND STANDARD DEVIATION OF POSE ERROR PER TEAT

### C. Pose Estimation Rate

Both the Detectron and Matterport implementation of MaskRCNN achieve a similar inference time (on 640x480 images) of roughly 150 ms on a Tesla T4 GPU on a remote server. We are confident that further engineering of the implementation could sensibly reduce it. Inference performance could be trivially increased by reducing processed image resolution (at the cost of lower mask precision). We still need to evaluate performance of the network on embedded GPU boards such as the Nvidia Jetson. Adding an estimated latency of 50ms per submitted image over a remote connection (e.g., with 5G) even with this initial setup would yield a processing rate of 5 FPS which is sufficient for limited teat movement.

The weakest part of our current algorithm implementation is in the transformation from the 2D teat masks to the corresponding 3D points in the point cloud. The current (trivial) implementation converts each point in the contour of a mask into a 3D ray to build a (pixel-precision) frustum. This operation, calculated for each point, is currently executed sequentially resulting in an average execution time of up to 50 ms. Reducing the number of considered contour points (e.g., sampling every ten pixels) can sensibly speed up the process with limited effect on the frustum precision.

A video of the overall system prototype in operation is visible online<sup>7</sup>.

## VII. CONCLUSIONS AND FUTURE WORK

This paper relates the initial work concerning the 3D sensor evaluation and teat pose estimation activities of a research project to build a next generation milking robot for the Swiss market. The current prototype already demonstrated sub-centimeter precision in teat pose estimation (albeit on a synthetic cow). The presented results are preliminary and will require further engineering and validation in real environments in subsequent steps of this and following projects.

There are several directions for the extension of this work. To improve the detection rate a faster segmentation network could significantly reduce the prediction time on RGB images.

Rather than relying on a sequence of arm movements based on estimated teat poses to approach the teats, an

active vision system could take into account occlusions, and learn the optimal sequence of positions to perform teat pose estimation for any shape of udder. We started working on such an “embodied AI” approach already in [19] and more work is ongoing.

## REFERENCES

- [1] B. R., K. S. R., and H. H., “The profitability of automatic milking on dutch dairy farms,” *Journal of Dairy Science*, vol. 90, no. 1, 2007.
- [2] K. W., H. A., and F. R., “Design of the robot arm of the astronaut a3 milking system,” *5th Int. Fluid Power Conference, Aachen*, pp. 353–364, 2006.
- [3] H. T., A. H., and B. E., “Zur ansetzgenauigkeit des melkzeuges bei ams,” *LANDTECHNIK*, vol. 3, 1999.
- [4] A. Pal, A. Rastogi, S. Myongseok, and B. S. Ryuh, “Algorithm design for teat detection system methodology using tof, rgbd and thermal imaging in next generation milking robot system,” in *2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*. IEEE, 2017, pp. 895–896.
- [5] M. Akhloufi, “3d vision system for intelligent milking robot automation,” in *Intelligent Robots and Computer Vision XXXI: Algorithms and Techniques*, vol. 9025. International Society for Optics and Photonics, 2014, p. 90250N.
- [6] A. Rastogi and B. S. Ryuh, “Teat detection algorithm: Yolo vs. haar-cascade,” *Journal of Mechanical Science and Technology*, vol. 33, no. 4, pp. 1869–1874, 2019.
- [7] M. van der Zwan and A. Telea, “Robust and fast teat detection and tracking in low-resolution videos for automatic milking devices,” in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3. SciTePress, 2015, pp. 520–530.
- [8] N. O’Mahony, S. Campbell, A. Carvalho, L. Krpalkova, D. Riordan, and J. Walsh, “3D Vision for Precision Dairy Farming,” *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 312–317, 2019.
- [9] A. Dorokhov, V. Kirsanov, D. Pavkin, S. Yurochka, and F. Vladimirov, “Recognition of Cow Teats Using the 3D-ToF Camera When Milking in the “Herringbone” Milking Parlor,” *Advances in Intelligent Systems and Computing*, vol. 1072, pp. 128–137, 2020.
- [10] G. Singh, S. Miao, S. Shi, and P. Chiang, “Fotonnet: A hw-efficient object detection system using 3d-depth segmentation and 2d-dnn classifier,” *arXiv preprint arXiv:1811.07493*, 2018.
- [11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [12] X. Han, H. Laga, and M. Bennamoun, “Image-based 3d object reconstruction: State-of-the-art and trends in the deep learning era,” *IEEE transactions on pattern analysis and machine intelligence*, 2019.
- [13] M. Ciocarlie, K. Hsiao, E. G. Jones, S. Chitta, R. B. Rusu, and I. A. Şucan, “Towards reliable grasping and manipulation in household environments,” in *Experimental Robotics*. Springer, 2014, pp. 241–252.
- [14] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, “Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics,” *arXiv preprint arXiv:1703.09312*, 2017.
- [15] A. ten Pas, M. Gualtieri, K. Saenko, and R. P. Jr., “Grasp pose detection in point clouds,” *CoRR*, vol. abs/1706.09911, 2017. [Online]. Available: <http://arxiv.org/abs/1706.09911>
- [16] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, “kpam: Keypoint affordances for category-level robotic manipulation,” *arXiv preprint arXiv:1903.06684*, 2019.
- [17] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, “Frustum PointNets for 3D Object Detection from RGB-D Data,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 918–927, 2018.
- [18] S. Cortesi, “Evaluation von 3d kameras,” *Institut für Mechatronische Systeme (IMS)*, Tech. Rep., 06 2019.
- [19] D. Roost, R. Meier, G. Toffetti Carughi, and T. Stadelmann, “Combining reinforcement learning with supervised deep learning for neural active scene understanding,” in *Active Vision and Perception in Human (-Robot) Collaboration Workshop at IEEE RO-MAN 2020 (AVHRC’20)*. University of Essex, 2020.

<sup>7</sup><https://www.youtube.com/watch?v=-7NiKSdA4AM>